Simple4All proposals for the Albayzin Evaluations in Speech Synthesis

Jaime Lorenzo-Trueba, Oliver Watts, Roberto Barra-Chicote, Junichi Yamagishi, Simon King, and Juan M Montero. Simple4all proposals for the albayzin evaluations in speech synthesis. In Proc. Iberspeech 2012, 2012.

Simple4All proposals for the Albayzin Evaluations in Speech Synthesis

Jaime Lorenzo-Trueba[1], Oliver Watts[2], Roberto Barra-Chicote[1], Junichi Yamagishi[2], Simon King[2], and Juan M. Montero[1]

[1]Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain
[2]CSTR, University of Edinburgh, United Kingdom
jaime.lorenzo@die.upm.es

Abstract. Simple4All is a European funded project that aims to streamline the production of multilanguage expressive synthetic voices by means of unsupervised data extraction techniques, allowing the automatic process of freely available data into flexible task specific voices. In this paper we describe three different approaches for this task, the first two covering enhancements in expressivity and flexibility with the final one focusing on the development of unsupervised voices. The first technique introduces the principle of speaker adaptation from average models consisting of multiple voices, with the second being an extension of this adaptation concept into allowing the control of the expressive strength of the synthetic voice. Finally, an unsupervised approach to synthesis capable of learning from unlabelled text data is introduced in detail.

Keywords: Emotional Speech Synthesis, Unsupervised Synthesis, Emotional Strength Control

1 Introduction

One of the goals in the Simple4all[1] project is the automatic modification of the speaking style of a neutral or expressive voice without requiring the recording of additional data. It should also be able to maintain the synthesized voice quality of the source voice but at the same time providing high recognition rates of the target expressive style. Additionally, because expressivity in real life is not a discrete space but a continuous space it is also required for the system to be able to mimic this property. Another aim of the project is to be able to generate all these voices with unlabelled data, as that significantly increases the potential sources of training data from which to produce the models.

This paper proposes three different systems that aim to fulfil different aspects of the project. The first and second systems (systems A and B) tackle the problem of enhancing and modifying the expressivity of voice models with as little data as possible by making use of adaptation techniques. The third system

[1] www.simple4all.org/

(system C) provides a state of the art system capable of training unsupervised synthetic voices.

Albayzin 2012, then, proves to be the perfect challenge for testing our systems as its main focus on enhancing the naturalness and emotional strength and identification rates while keeping the similarity with the source's natural voice is the definite proving ground for the prosed systems.

Regarding the structure of the present paper, Sections 2, 3 and 4 describe the three proposed systems, namely: Average Model based Voices, Adaptation-based Emotional Strength Controlled Voices and the Unsupervised Front-End in that order. Section 3 also includes a short review on a previous perceptual analysis of the system to justify its inclusion. Finally, section 5 includes the results of the challenge and conclusions to be drawn from them.

2 System C: Average Model based Voices

Because sometimes there is only a small amount of data available for training the voices, or even because the processing time of the training is critical, systems based in the creation of average models and obtaining adapted voices from them such as System **C** have been created. This particular system is based on obtaining the acoustic models of the different emotions of the speakers through a model-space SAT algorithm [1] to then proceed to use a shared decision-tree based clustering algorithm such as [2] in order to tie the parameters and define the speaker average model. From this average voice that encompasses all the different versions of the speaker voice (neutral, happy, sad, angry and surprised in this evaluation), the particular emotional model is then adapted using the CSMAPLR algorithm [3] (Figure 1).

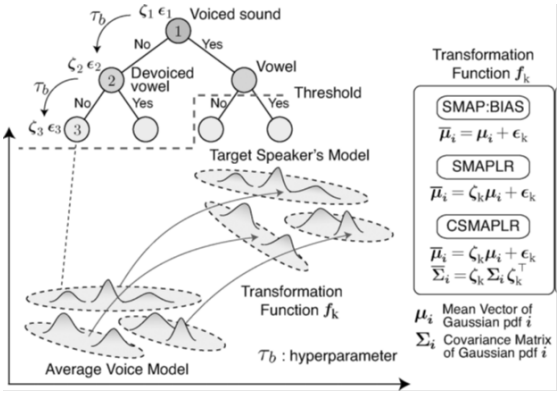

Fig. 1. Schematic defining the CSMAPLR average voice generation process.

The voices obtained in this fashion are much more robust in situations in which the training data is limited, as it pools all the available resources in order

to attain stability, and then modifies the parameters so that the obtained model resembles more closely the intended voice.

3 System D: Adaptation-based Emotional Strength Controlled Voices

It is known that HMM-based modelling introduces a smoothing in the synthetic voices, reducing the expressive capabilities of the model. This side effect is further enhanced by the adaptation process. Consequently, in this particular task where it is important to synthesize expressive voices it becomes necessary to find a way to enhance said expressivity. With that purpose in mind we developed system **D**. In this system, the adapted models of the different emotions are obtained from the average voice in the same fashion as in system **C**, but they are only used to obtain the transformation function capable of morphing the neutral model into the different emotional models (Figure 2). Then, through a control ratio it becomes possible to either enhance or attenuate the expressivity of the transformed model, allowing for a continuous modeling of the emotional space.

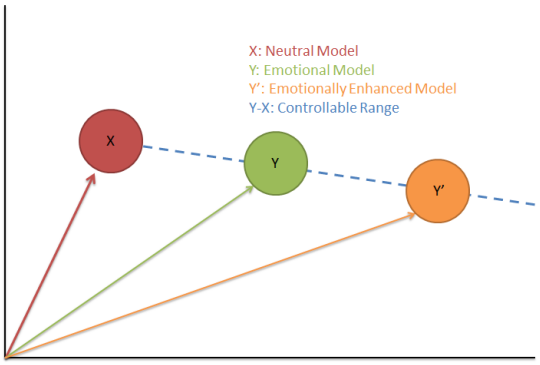

Fig. 2. Representation of the emotional strength control system.

3.1 Analysis of the ES-Control effects

This system has been studied through a perceptual test, which analysed the emotion identification rates, perceived naturalness and perceived emotional strength for the adaptation-based system with control ratios ranging from 0.0 to 2.0, always comparing it with its natural voice.

Regarding identification rates, for control ratios of 1.5 times the plain adapted emotion and upwards the results become comparable to those of natural voice, sometimes even surpassing them. This trend has been proven to be stable at

least up to control ratios of 2.0, but it would be hard to extrapolate them further than that due to natural speech rarely being so full blown emotional. For naturalness, considering the natural voice obtained a 4 out of 5 in the likert scale, the ES-controlled voice manages to attain constant values of 3, stable in all the considered ratios, which proves the usefulness of enhancing the expressivity in this fashion as it does not alter the attained speech quality. Finally, the perceived emotional strength was seen to be practically linear throughout the examined control ratios range, corroborating the hypothesis that a continuous emotional strength space can be attained. Additionally, it was also since control ratios between 1.5 and 2.0 that the perceived strength results matched those of natural voice, once again sometimes surpassing them.

All in all, the conclusion of this study was that applying emotional strength control through adaptation can provide significant enhancements to the perceivable expressivity of the synthetic voices without incurring in naturalness and speech quality penalties for control ratios of at least up to 2.0, with possibly 1.5 being the optimal value that does not suffer from over-exaggeration effects in the synthesized voice.

4 System E: Unsupervised Front-End

System **E** was built to test the prototype text-processing modules which are being developed as part of the *Simple4All* project. The aim of developing these modules is to provide a TTS front-end which makes few implicit assumptions about the target language, and which can be configured with minimal effort and expert knowledge to suit arbitrary new target languages. To this end, the modules rely on resources which are intended to be universal, such as the Unicode character database[2], and make use of unsupervised learning so that unlabelled text resources can be exploited without the need for costly annotation. The initial version of this front-end is based very closely on the ideas outlined in [4], and is a re-implementation with some modifications of the system that was there used to build synthetic voices in English, Romanian and Finnish. The version of the front-end used in the current Challenge will be briefly described here, as well as the training of acoustic models using the annotation provided by it.

4.1 Text Analysis

Data The ISO 8859-1 text of the transcriptions provided for the Challenge was first manually converted to UTF-8 encoding, which is required by the front-end. After this initial conversion, however, text processing is fully automatic as described below. In addition to the text transcription of the speech, we used an additional 1.4 million words of untranscribed Spanish text. This text was selected to roughly match the domain of the speech corpus: the entire text of Cervantes' *Don Quijote* (c. 400,000 words) was taken from Project Gutenberg[3] to match

[2] `www.unicode.org/ucd/`
[3] `www.gutenberg.org`

the Quijote portion of the corpus, and the text of c. 1800 news stories drawn randomly from those published by the *El Mundo* newspaper in 2006 (giving c. 1 million words) was used to match the remaining part of the corpus

Tokenisation All language-specific knowledge used to provide the chunking of text into words and punctuation symbols is derived from the Unicode character database. Each input character is assigned a *coarse category* by lookup in the that database. The three coarse categories used are formed by merging Unicode general categories: coarse category *letter* is made by made by combining general categories L and N, coarse category *space* maps exactly to general category Z, and all remaining general categories map to a coarse category called *punctuation*. Tokenisation is performed by placing token delimiters at the ends of contiguous sequences of characters belonging to the coarse category *letter*. The chunks between these delimiters are tokens. For example (if whitespace is represented for clarity by an empty box) the training utterance:

sí.□con□seguridad.

is chunked into the following six tokens:

sí .□ con □ seguridad .

Furthermore, the coarse categories are used to assign each token a token class, from an inventory of three classes that share the names of the coarse categories. The coarse categories are placed in the following order of precedence: letter, punctuation, space. A given token is classified by traversing this sequence of categories from left to right; if all characters in that token belong to the current category or to a category on the left, the token is assigned to that token class. The above chunks are assigned the following token classes:

letter punct. letter space letter punct.

For implementational reasons (e.g. to enable use of tools which can only handle ASCII characters and to avoid confusion with special symbols such as field delimiters during processing), each input character is assigned a *safetext* form, consisting of a string of characters belonging to the 52 upper- and lower-case letters of the English alphabet. Safetexts consisting of multiple characters but corresponding to a single character of surface text are delimited by underscores to allow unambiguous mapping back to surface forms as needed. Where a surface form is not already a safetext, one is constructed automatically from the name of the character in the Unicode database. The 6 chunks above are rewritten in safetext form as follows:

s_LATINSMALLLETTERIWITHACUTE_
_FULLSTOP__SPACE_
con
SPACE
seguridad
FULLSTOP

All characters encountered in the text of the training data are stored in a text file along with their automatically generated coarse categories and safetexts. This allows a user to manually intervene to correct bad categorisation of characters due to text encoding mistakes or due to errors in the Unicode database, and to specify more user-friendly safetexts than the automatically generated ones (see e.g. the long-winded *_LATINSMALLLETTERIWITHACUTE_* for the character *í* in the example above). For building the voices presented here, however, automatically generated categories and safetexts were used with no manual intervention.

Note that the present front-end requires text to be expanded fully before it is input to the system, and does not handle numerals and abbreviations. Correctly handling such non-standard words in a way that requires only minimal expert supervision is a topic of ongoing research [5].

Naive 'lexicon' and time alignment A naive 'lexicon' is used, in which the safetexts of letters of 'letter'-class tokens are used directly as the names of speech modelling units, in place of the phonemes of a conventional front-end. This has given good results for languages with transparent alphabetic orthographies such as Romanian, Spanish and Finnish, and can give acceptable results even for languages with less transparent orthographies, such as English [4, 6–8]. Using the pronunciations provided by this lexicon, a set of labels is initialised for the speech part of the database by iteratively estimating a set of HMMs and using these to force-align the speech with the labels using a procedure based very closely on that described in [9]. Tokens which are assigned by the tokeniser to the *punctuation* and *space* token classes are allowed by the naive lexicon to be pronounced both as a silence symbol (*sil*) or as a non-emitting symbol (*skip*). As well as determining the timing of letter-boundaries, therefore, the forced alignment procedure determines which space and punctuation tokens are realised as a pause, and which are skipped.

Letter- and word-representations The system makes use of no expert-specified categories of letter and word, such as phonetic categories (vowel, nasal, approximant, etc.) and part of speech categories (noun, verb, adjective, etc.). Instead, features that are designed to stand in for such expert knowledge but which are derived fully automatically from the distributional analysis of the text corpus are used. The distributional analysis is conducted via vector space models (VSMs); the VSM was applied in its original formulation as a model for Information Retrieval to the characterisation of documents. VSMs are applied to TTS in [4], where models are built at various levels of analysis (letter, word and utterance) from large bodies of unlabelled text. To build these models, co-occurrence statistics are gathered in matrix form to produce high-dimensional representations of the distributional behaviour of e.g. word and letter types in the corpus. Lower-dimensional representations are obtained by approximately factorising the matrix of raw co-occurrence counts by the application of slim singular value decomposition. This distributional analysis places textual objects

in a continuous-valued space, which is then partitioned during the training of TTS system components such as acoustic models for synthesis or decision trees for pause prediction. For the present voices, a VSM of letters was constructed by producing a matrix of counts of immediate left and right co-occurrences of each letter type, and from this matrix a 5-dimensional space was produced to characterise letters. Token co-occurrence was counted with the nearest left and right neighbour tokens which are *not* of the class *space*; co-occurrence was counted with the most frequent 250 tokens in the corpus. A 10-dimensional space was produced to characterise tokens.[4]

Pause prediction The system uses a decision tree to predict whether a token of class *space* or *punctuation* is realised as a pause or not. Data for training the tree is produced from the time-aligned transcriptions of the training data. The predictor variables used for tree training are the token class of the token in question (i.e. whether it is punctuation or space) and the VSM features of the tokens preceding and following the token. The annotation of training data is done by detection of silence in the audio during forced alignment as already described. At run-time, the tree's predictions are used.

Rich contexts Information extracted from the utterance structures resulting from text processing is used to create a set of rich contexts for each speech unit in the database. Features include the identity of the letter to be modelled and those of its neighbours (within a 5-letter window), the VSM values of each of those letters, and the distance from and until a word boundary, pause, and utterance boundary. Word VSM features were not included directly in the contexts, but were used by the decision tree for predicting pauses at runtime. It should be emphasised that there is nothing language-specific about these features: they are generally applicable to any language making use of an alphabetic script and marking word boundaries orthographically.

4.2 Acoustic Models

Emotion-dependent acoustic models were built for each of the angry, happy, neutral and surprised subsets of the database, using the distributed 48kH waveforms and the acoustic parameters described in [11], and using a now standard speaker dependent recipe described in [12]. The *sad* subset of the data created problems for the extraction of STRAIGHT features, and so the least problematic part of that subset (the *Quijote* section) was used to adapt the acoustic models built for the neutral condition to the sad condition (using a combination of the CSMAPLR adaptation and MAP adaptation techniques, as in [13]), although it was found that using the unadapted neutral duration model gave best results. The waveforms of the synthesised test-set were downsampled to the required 16kH for submission.

[4] The package *Gensim* [10] was used for performing the singular value decomposition needed to obtain these features.

5 Results

The first conclusion that can be extracted is that even if synthetic voices do not present the same quality results as the natural voice yet, although for emotion identification rates, emotional strength and speaker similarity the obtained results show that the presented systems can compare and sometimes even surpass the recognition potential of the natural voice. This is a feat that encourages us for developing even further the systems in order to one day obtain the most natural possible voices. In a system-by-system analysis of the results, system **C**,

SYSTEM	Speech Quality	Emotional Strength	Speaker Similarity	Emotion Identification Rate	Relative Performance
A	0.87	0.71	0.43	0.78	**1.00**
B	0.44	0.41	0.46	0.53	**0.32**
C	0.57	0.53	0.42	0.65	**0.51**
anger	0.62	0.64	0.37	0.78	0.51
happiness	0.57	0.61	0.48	0.65	0.80
neutral	0.51	0.33	0.47	0.61	0.60
sadness	0.49	0.52	0.41	0.65	0.48
surprise	0.65	0.57	0.38	0.50	0.29
D	0.53	0.54	0.39	0.63	**0.46**
anger	0.61	0.67	0.36	0.82	0.53
happiness	0.53	0.59	0.40	0.34	0.38
neutral	0.55	0.31	0.36	0.76	0.58
sadness	0.49	0.56	0.32	0.79	0.47
surprise	0.51	0.59	0.51	0.37	0.25
E	0.41	0.43	0.42	0.46	**0.27**
anger	0.35	0.38	0.46	0.30	0.13
happiness	0.44	0.53	0.35	0.46	0.36
neutral	0.43	0.32	0.34	0.74	0.46
sadness	0.43	0.43	0.43	0.38	0.26
surprise	0.42	0.51	0.55	0.38	0.21

Table 1. Normalized results for the proposed systems [14]. All four measures take values from 0 to 1, with the performance being obtained as $4 * (SQ * ES * SS * EIR)/(SQ + ES + SS + EIR)$ and then normalized by the natural speech (system **A**).

is clearly the one with the best overall results, constantly placing first or second in all the categories excepting speaker similarity. This results prove the usefulness of using averaged data and then adapting into the particular emotion in this situation in which the training data is not very extensive. One of the best conclusions that can be extracted from the results of this system is that they are significantly stable, not showing any kind of unexplainable dip in any emotion or measured category. The only exception would be the recognition rate of surprise, which can be justified by the significant confusion it presented with happiness.

System **D** shows how the strength control process is promising, although it can be improved. The expected results would be to obtain the better identification rates and emotional strength at the cost of speech quality, and so it is but with some significant constraints. First of all, the confusion between happiness and surprise was greatly enhanced to the point where both emotions become almost unrecognisable between each other. This is thought to be because all the different features were extrapolated, and it is expected that extrapolating only some particular features (such as modifying the spectral parameters while leaving the F0 intact) would help solve this problem. Emotional strength results show that they are marginally higher or lower than those of system **C**, but not in a statistically relevant way. It is also expected that applying the partial control would help relief this perception problems.

Finally, the unsupervised system (system **E**), shows some very promising results in the sense that given the handicap of being unsupervised, all the different measures are comparable with other supervised systems in the competition, sometimes even surpassing it in speaker similarity. The biggest problem is presented when considering the emotional measures was that the voices obtained through this system, while very clear and natural sounding, are sometimes too flat-sounding and neutral. One additional cause for this was that the feature extraction section of this system had some problems when extracting the fundamental frequency of the training data, which made the resulting voiced even more flat. Nevertheless the synthetic voice still keeps its natural properties as it can be seen in the speaker similarity measures, in which this system sometimes obtains the best results.

6 Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO(TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo has been funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

References

1. J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.
2. J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.* IEEE, 2007, vol. 4, pp. IV–1233.

3. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.
4. Oliver Watts, *Unsupervised Learning for Text-to-Speech Synthesis*, Ph.D. thesis, University of Edinburgh, 2012.
5. Ruben San-Segundo, Juan M. Montero, Veronica Lopez-Ludeña, and Simon King, "Detecting acronyms from capital letter sequences in Spanish," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 2012.
6. A. Black and A. Font Llitjos, "Unit selection without a phoneme set," in *IEEE TTS Workshop 2002*, 2002.
7. G.K. Anumanchipalli, K. Prahallad, and A.W. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 4645–4648.
8. Matthew P. Aylett, Simon King, and Junichi Yamagishi, "Speech synthesis without a phone inventory," in *Interspeech*, 2009, pp. 2087–2090.
9. Robert A. J. Clark, Korin Richmond, and Simon King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.
10. Radim Rehurek and Petr Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, May 2010, pp. 45–50, ELRA.
11. Junichi Yamagishi and Oliver Watts, "The CSTR EMIME HTS System for Blizzard Challenge," in *Proc. Blizzard Challenge 2010*, Sept. 2010.
12. H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.
13. Junichi Yamagishi, Heiga Zen, Tomoki Toda, and Keiichi Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, aug 2007.
14. Iberspeech2012, "2012 albayzin evaluations: speech-synthesis results," 2012.